Winning Space Race with Data Science

Mun Yin Ting
26 September 2021

Github: [Applied Data Science Capstone]

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Executive Summary

The Capstone project aims to discover the relationship between various variables with the success rate of a rocket launch. Various methodologies were carried out and results were concluded.

Methodologies:

1. Data Collection: Ready Data + Web Scraping
2. Data Preprocessing
3. Data Exploration: SQL + EDA + Dashboard
4. Location Analysis: Folium Map
5. Machine Learning PRediction

Results:

1. KSC LC-39A site has the highest success rate, with 77% of the launches performed are success
2. Payload Mass 9600 KG uses B4 Booster and has equal chance in fail and success
3. Several model can be used for prediction in success or failure in landing: SVM, Decision Tree, Logistic Regression, KNN

Introduction

Background:

SpaceX advertises Falcon 9 rocket at a very much lower cost compared to other providers due to its reusability of first stage. SpaceY as a competitor would like to know the cost of each launch and how to reduce it by analyzing Space X's data.

Problems:

First stage is the most expensive in the launch, if first stage can land successfully, it will cut down the cost.

So, how do we know if a first stage launch is successful?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Existing dataset from SpaceX and Web scraping from Wikipedia
- Perform data wrangling
  - Remove missing values, scaled, get dummies, categorise outcome to successful or fail and assign binary values
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Build and Tune using GridSearch, accuracy up to 83.3%

Data Collection

Existing Dataset from SpaceX API

Data from Wikipedia

Data Collection – SpaceX API

[GitHub click here]

Task 1: Request and parse the SpaceX launch data using the GET request

To make the requested JSON results more consistent, we will use the following static response object for this project:

```
In [9]: static_json_url='https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/API_call_spacex_api.json'
```

We should see that the request was successfull with the 200 status response code

```
In [10]: response.status_code
Out[10]: 200
```

Now we decode the response content as a Json using `.json()` and turn it into a Pandas dataframe using `.json_normalize()`

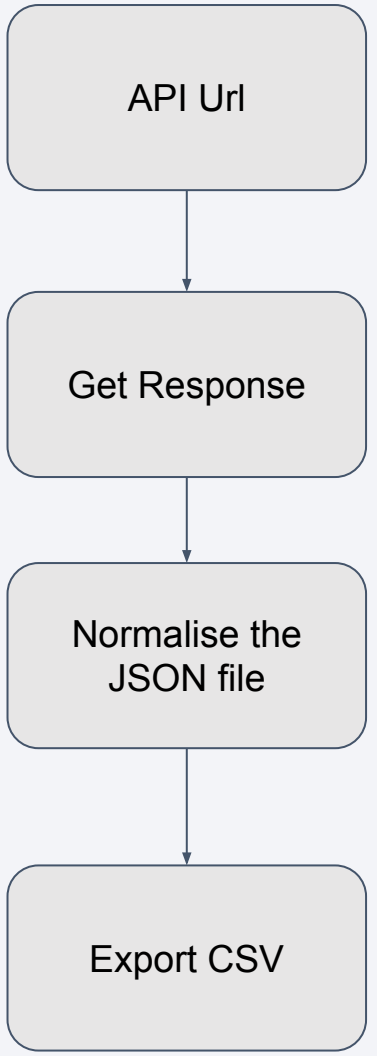
```
In [12]: # Use json_normalize meethod to convert the json result into a dataframe
response = requests.get(static_json_url)
response.json()
data = pd.json_normalize(response.json())
```

Using the dataframe `data` print the first 5 rows

```
In [13]: # Get the head of the dataframe
data.head()
```

Out[13]:

	static_fire_date_utc	static_fire_date_unix	tbd	net	window	rocket	success	details	crew	ships	capsules	payloads
0	2006-03-17T00:00:00.000Z	1.142554e+09	False	False	0.0	5e9d0d95eda69955f709d1eb	False	Engine failure at 33 seconds and loss of vehicle	[]	[]	[]	[5eb0e4b5b6c3bb0



Data Collection - Scraping

GitHub click here

Create a BeautifulSoup object from the HTML response

```
In [10]: # Use BeautifulSoup() to create a BeautifulSoup object from a response text content
         soup = BeautifulSoup(response, 'html5lib')
```

Print the page title to verify if the BeautifulSoup object was created properly

```
In [11]: # Use soup.title attribute
         soup.title
Out[11]: <title>List of Falcon 9 and Falcon Heavy launches - Wikipedia</title>
```

TASK 2: Extract all column/variable names from the HTML table header

Next, we want to collect all relevant column names from the HTML table header

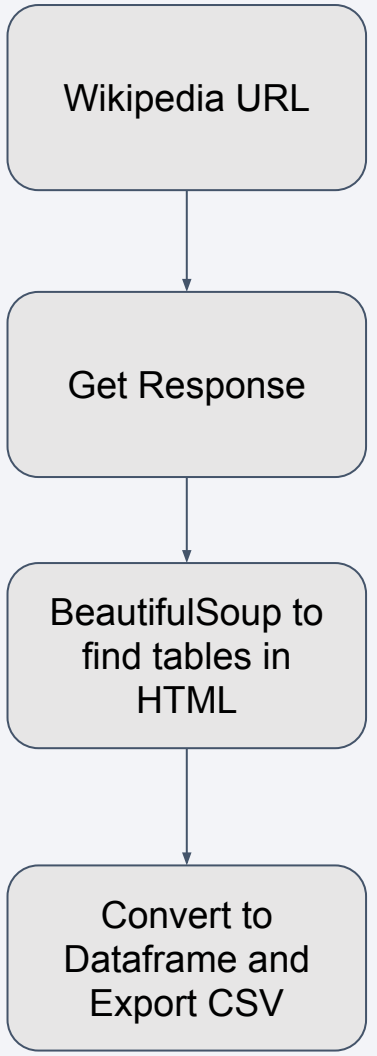
Let's try to find all tables on the wiki page first. If you need to refresh your memory about BeautifulSoup, please check the external reference link towards the end of this lab

```
In [13]: # Use the find_all function in the BeautifulSoup object, with element type `table`
         # Assign the result to a list called `html_tables`
         html_tables = soup.find_all('table')
```

Starting from the third table is our target table contains the actual launch records.

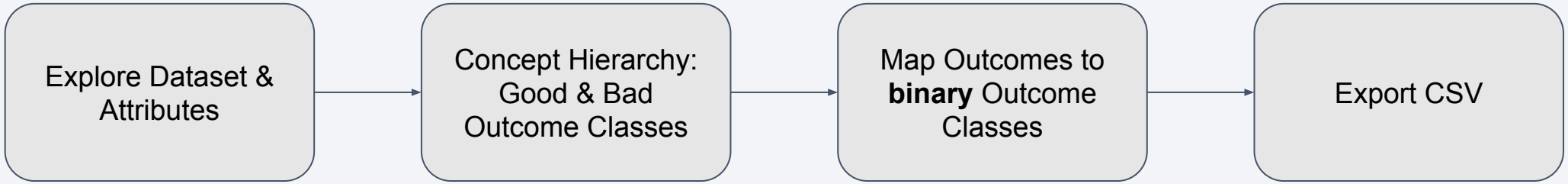
```
In [14]: # Let's print the third table and check its content
         first_launch_table = html_tables[2]
         print(first_launch_table)
```

```
<table class="wikitable plainrowheaders collapsible" style="width: 100%;">
<tbody><tr>
<th scope="col">Flight No.
</th>
<th scope="col">Date and<br/>time (<a href="/wiki/Coordinated_Universal_Time" title="Coordinated Universal Time">UTC</a>)
</th>
<th scope="col"><a href="/wiki/List_of_Falcon_9_first-stage_boosters" title="List of Falcon 9 first-stage boosters">Version,<b
r/>Booster</a> <sup class="reference" id="cite_ref-booster_11-0"><a href="#cite_note-booster-11">[b]</a></sup>
</th>
<th scope="col">Launch site
</th>
```



Data Wrangling

GitHub click here



The main step in Data Wrangling is Concept Hierarchy.

All outcomes that are ‘None’ or ‘False’ are concluded and generalised as ‘Bad outcome’. This generalises the outcome to a binary class, allows efficient and easy binary classification that has higher accuracy.

EDA with Data Visualization

GitHub click here

Chats plotted & discoveries:

1. Scatter plot (Payload mass vs Flight Number): increase proportionally (linear relationship), **more success** when number of flights and payload mass increases
2. Scatter plot (Launch site vs Flight Number): CCAFS SLC 40 has most flights and most success. Flight number increase, number of success generally increases
3. Scatter plot (Payload mass vs Launch Site): VAFB SLC 4E generally focuses on lower to middle payload mass with high success rate
4. Bar chart (Orbit vs Success Rate): ESL1, GEO, HEO, and SSO orbits has the highest success rate with an impressive 100% success
5. **Line (Success vs Year): Success rate has increased over the year! Technology has developed rapidly with exponential increase!**

EDA with SQL

GitHub click here

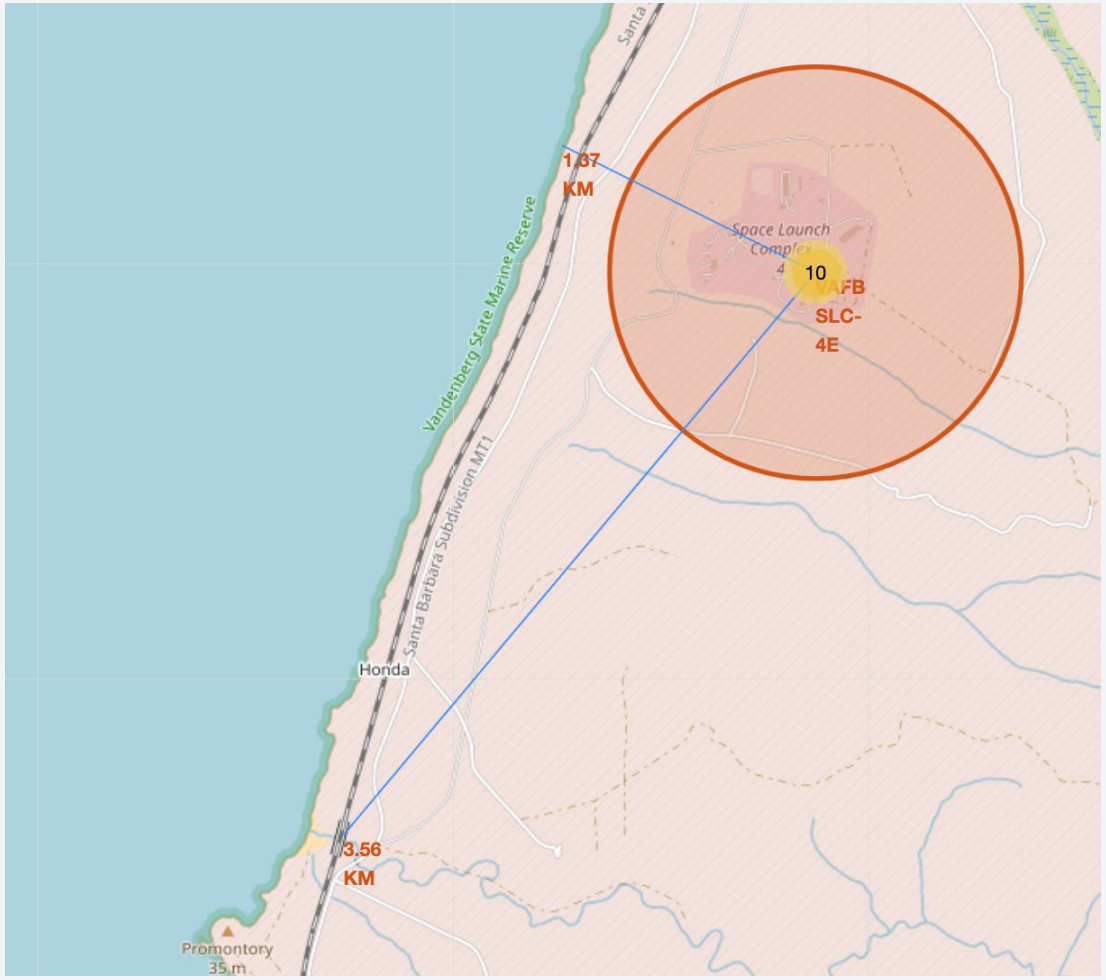
- Find launch sites involved
- Find launches that are located in CCA
- Find total payload mass carried by boosters launched by NASA (CRS)
- Find average payload mass carried by booster version F9 v1.1
- Find the date of first success landing
- Find all boosters that brought success landing with payload mass in between 4000 and 6000 kg
- Find total success and failure mission outcomes
- Find booster versions that carried maximum payload mass
- Find booster versions and launch site that failed in mission
- Find ranking of all outcomes in terms of frequency

Build an Interactive Map with Folium

GitHub click here

Object used:

- markers (to mark location/objects)
- circles (to highlight area in a radius)
- lines (to show distance between two locations)
- clusters (to show number of clusters, and how many child the cluster has)

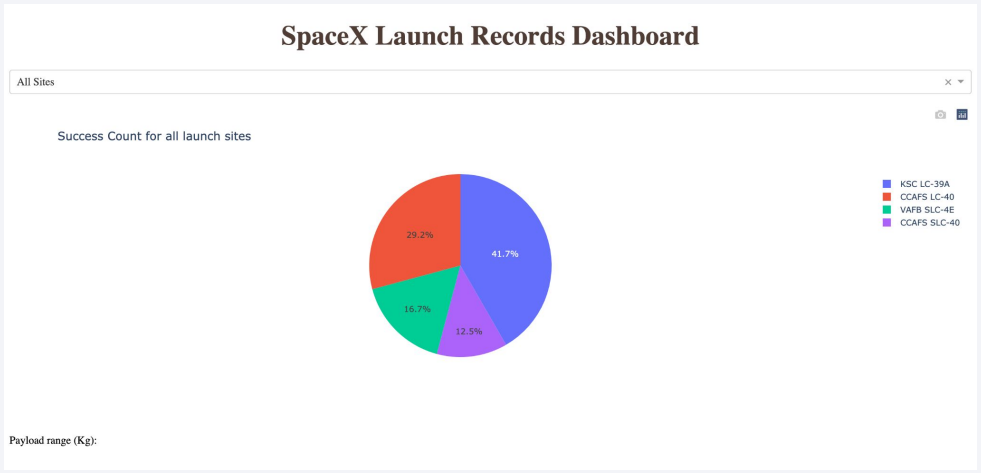


Build a Dashboard with Plotly Dash

GitHub click here

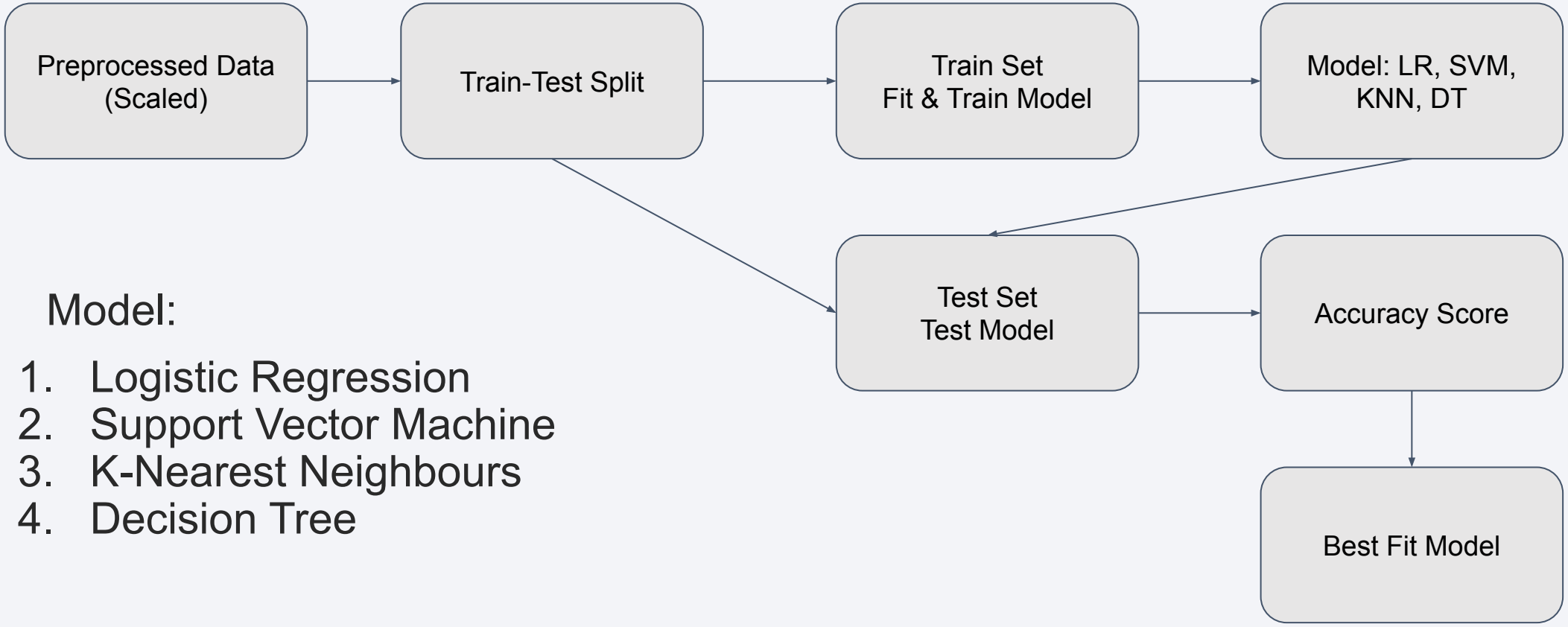
- Plots & Graphs: Pie chart & Scatter Plot
- Both visualizations are interactive to selected launch site and range of payload mass

- Pie chart: Success rate at all site/selected sites
- Scatter plot: Booster version in success and failure launches based on payload mass



Predictive Analysis (Classification)

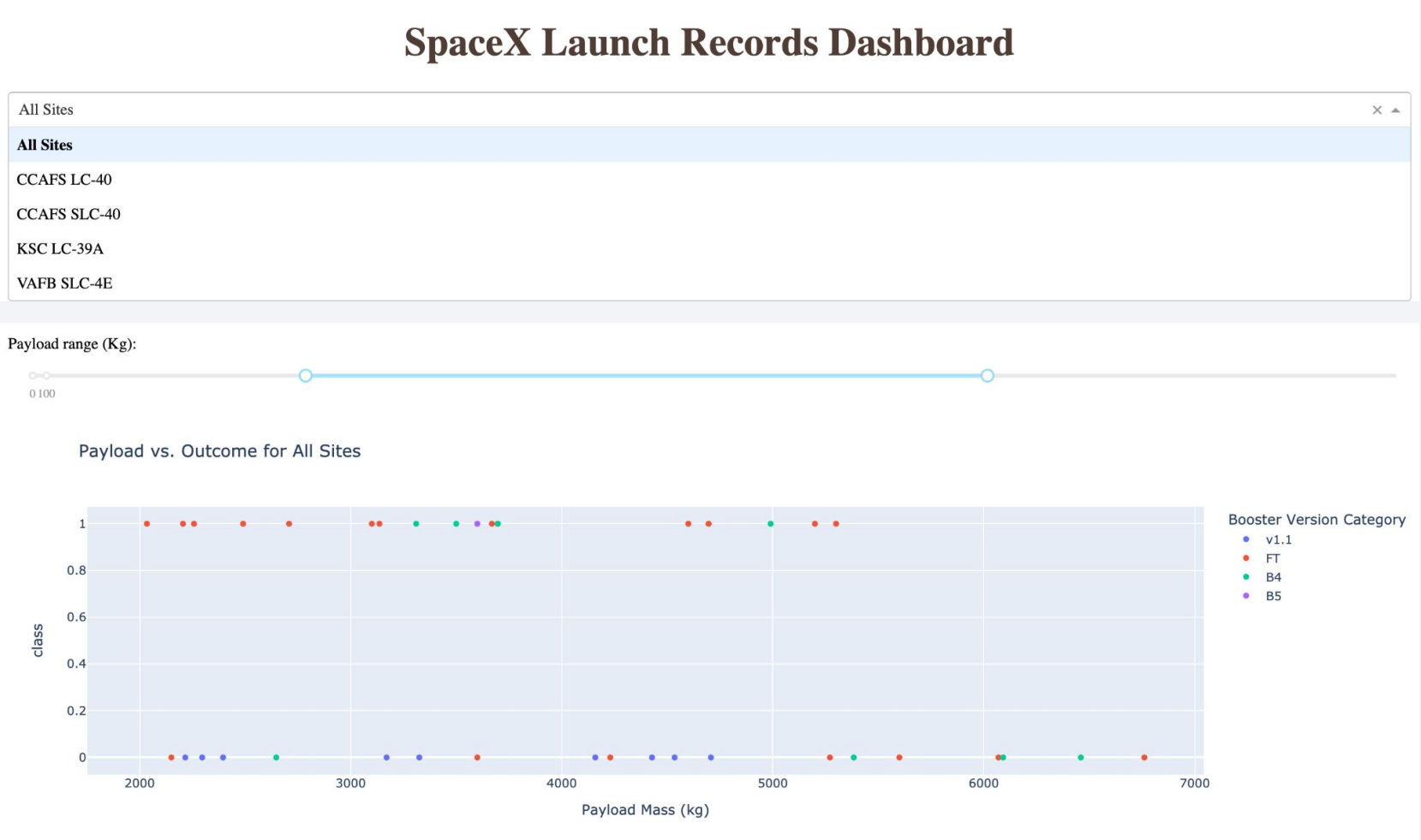
GitHub click here

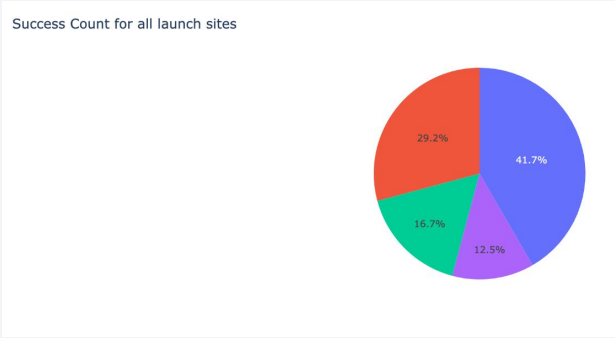


Model:

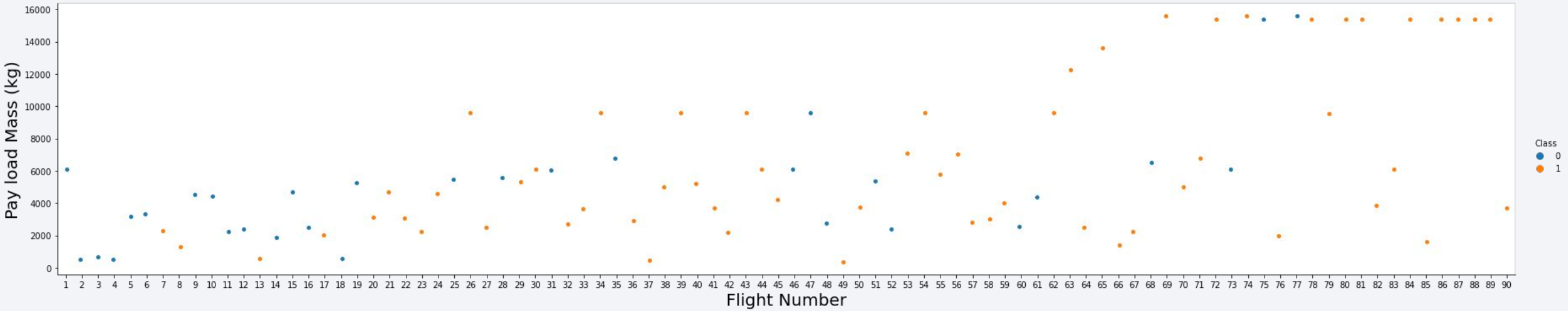
1. Logistic Regression
2. Support Vector Machine
3. K-Nearest Neighbours
4. Decision Tree

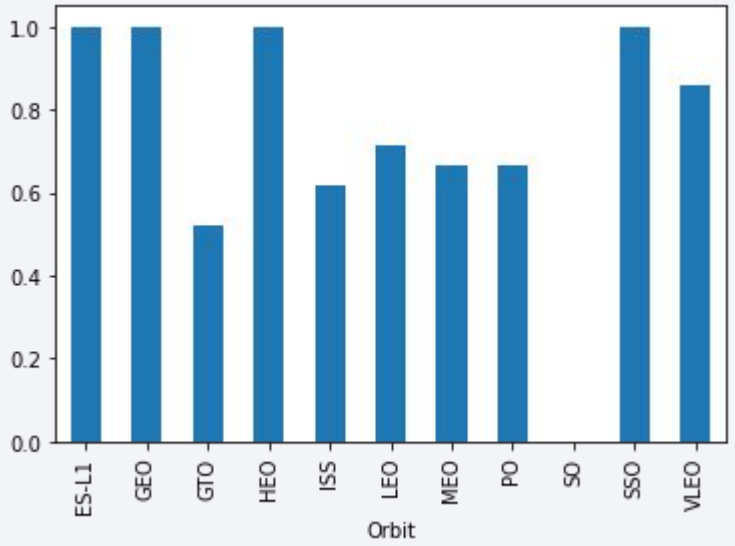
Results

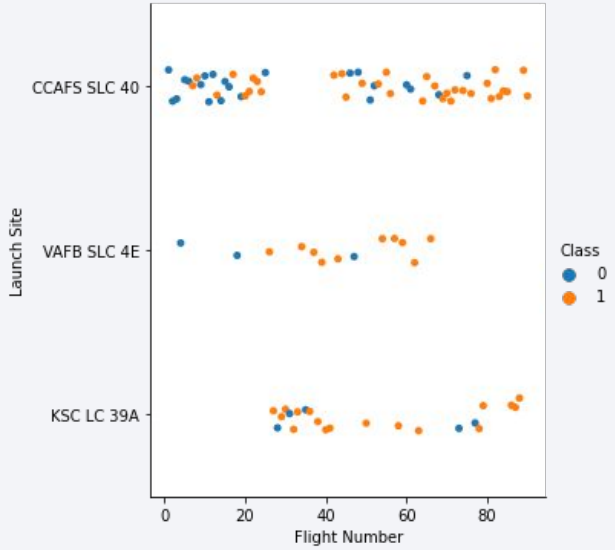

SpaceX Launch Records Dashboard
All Sites
All Sites
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E
Payload range (Kg):
0 100
Payload vs. Outcome for All Sites
class
Payload Mass (kg)
Booster Version Category
v1.1
FT
B4
B5


Success Count for all launch sites
29.2%
41.7%
16.7%
12.5%

Results


Pay load Mass (kg)
Flight Number
Class
0
1


ES-L1
GEO
GTO
HEO
ISS
LEO
MEO
PO
SO
SSO
VLEO
Orbit


CCAFS SLC 40
VAFB SLC 4E
KSC LC 39A
Launch Site
Flight Number
Class
0
1


Class

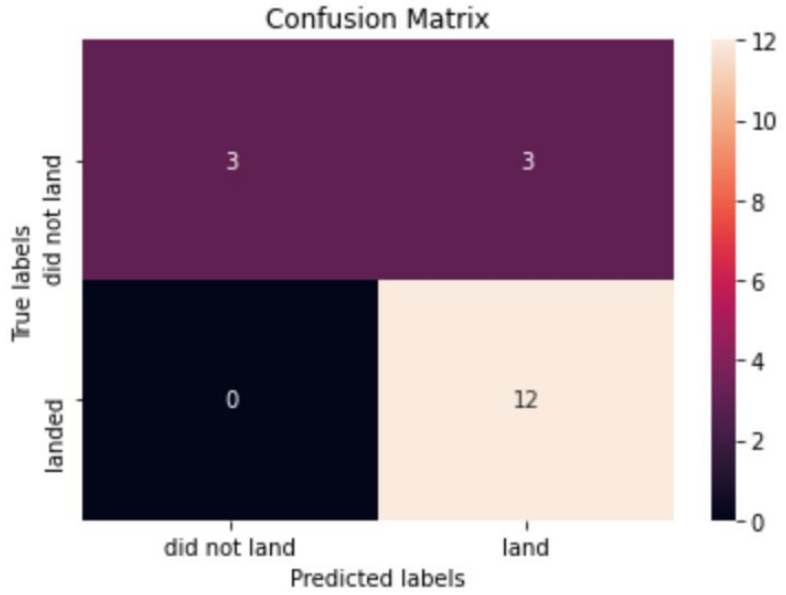
Results

```
logreg_cv.score(X_test, y_test)
```

```
0.8333333333333334
```

Lets look at the confusion matrix:

```
yhat=logreg_cv.predict(X_test)
plot_confusion_matrix(y_test,yhat)
```

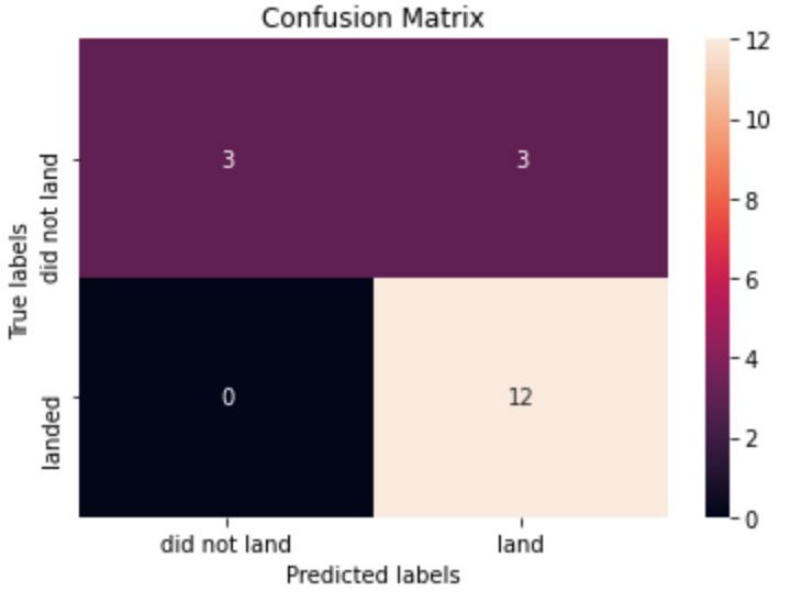



```
svm_cv.score(X_test, y_test)
```

```
0.8333333333333334
```

We can plot the confusion matrix

```
yhat1=svm_cv.predict(X_test)
plot_confusion_matrix(y_test,yhat1)
```



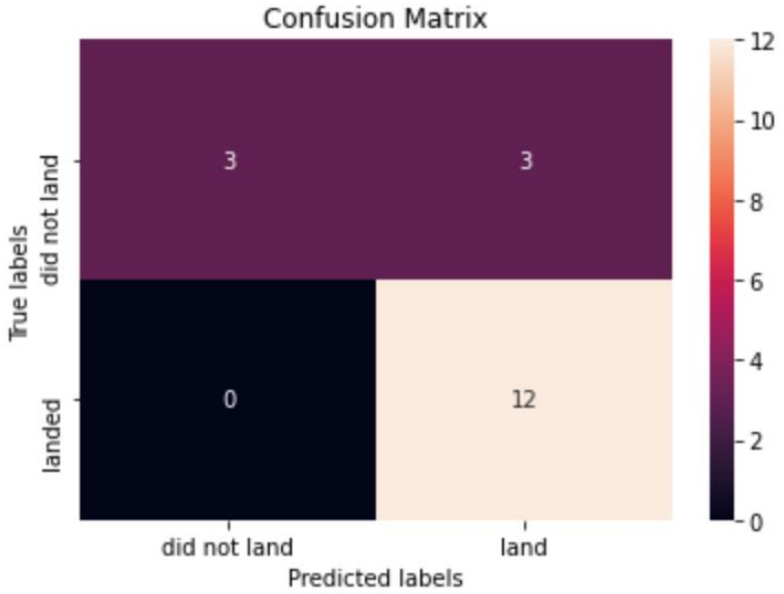
Results

```
tree_cv.score(X_test, y_test)
```

```
0.8333333333333334
```

We can plot the confusion matrix

```
yhat2 = tree_cv.predict(X_test)
plot_confusion_matrix(y_test,yhat2)
```

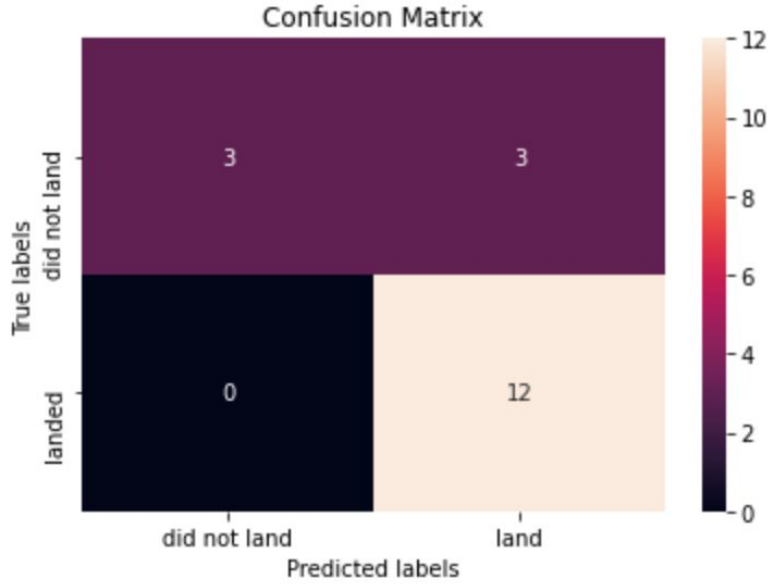


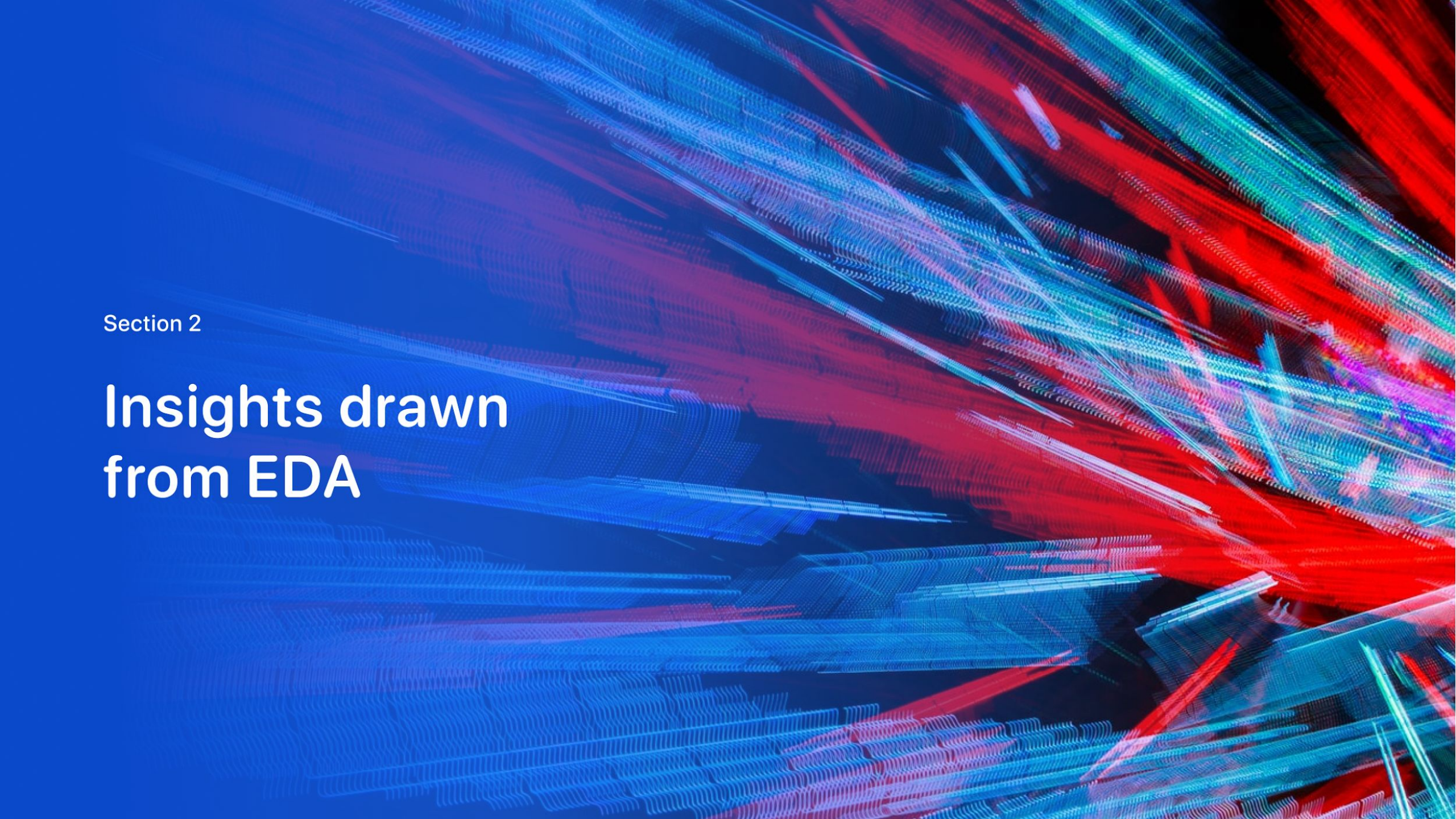

```
knn_cv.score(X_test, y_test)
```

```
0.8333333333333334
```

We can plot the confusion matrix

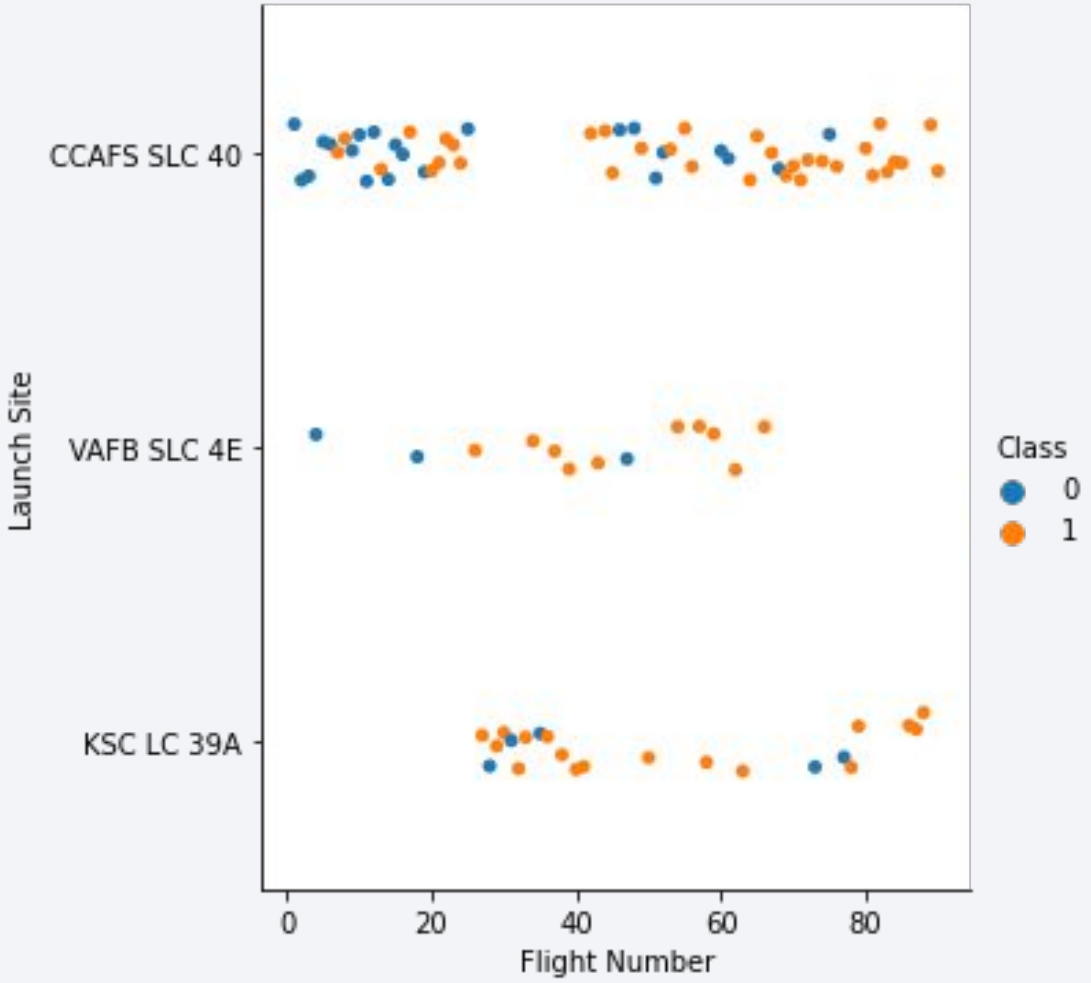
```
yhat3 = knn_cv.predict(X_test)
plot_confusion_matrix(y_test,yhat3)
```




Section 2
Insights drawn from EDA

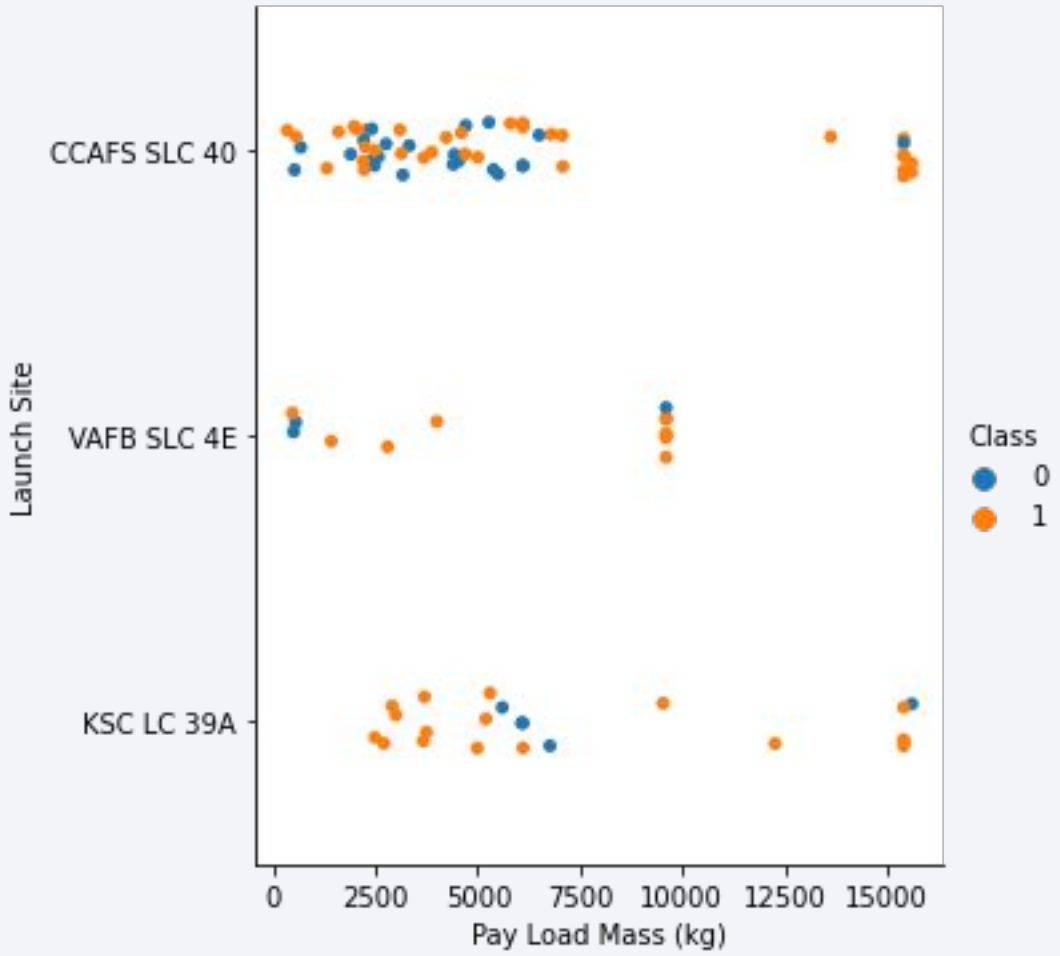
Flight Number vs. Launch Site

- CCAFS SLC 40 has the most flights/launches
- VAFB SLC 4E has lesser launched
- The more flights, the higher success rate



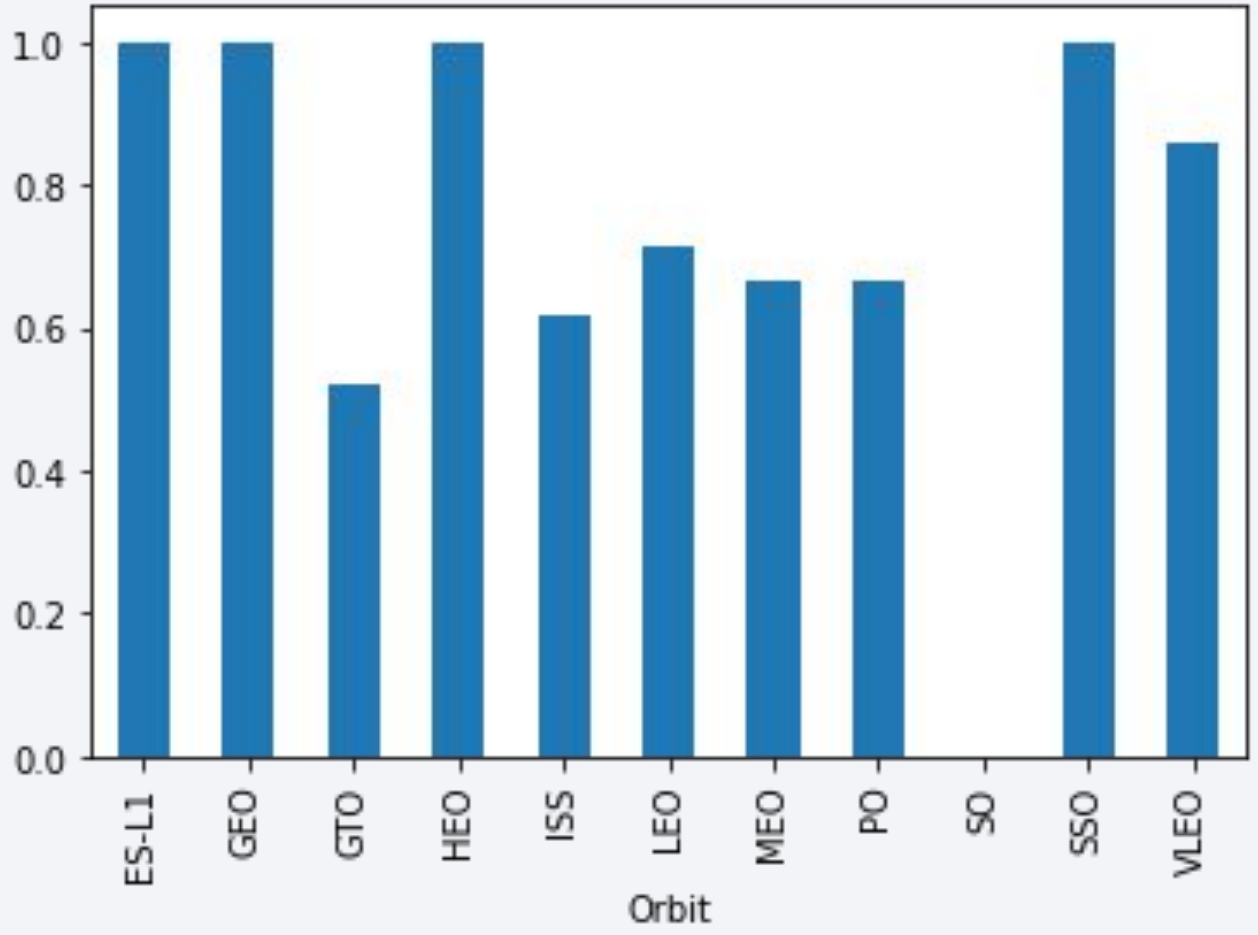
Payload vs. Launch Site

- VAFB SLC 4E only launched payload mass less than 10000 kg
- Higher payload mass has greater success rate
- Lower payload mass (lesser than 5000 kg) do not launch well at KSC LC 39A



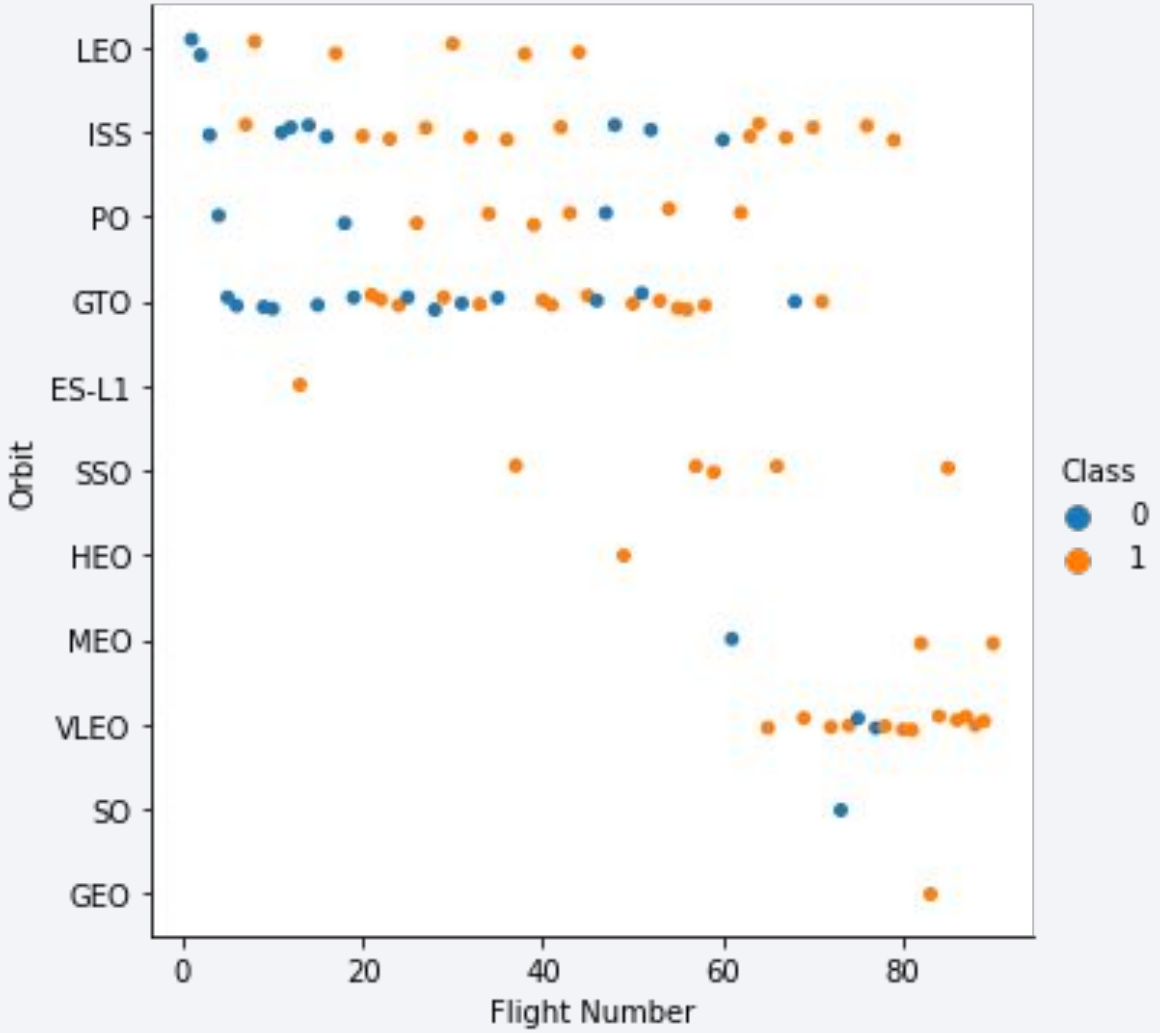
Success Rate vs. Orbit Type

- Orbit ES-L1 GEO, HEO and SSO has the highest success rate
- SpaceY can target to these orbits for higher success rate in launching



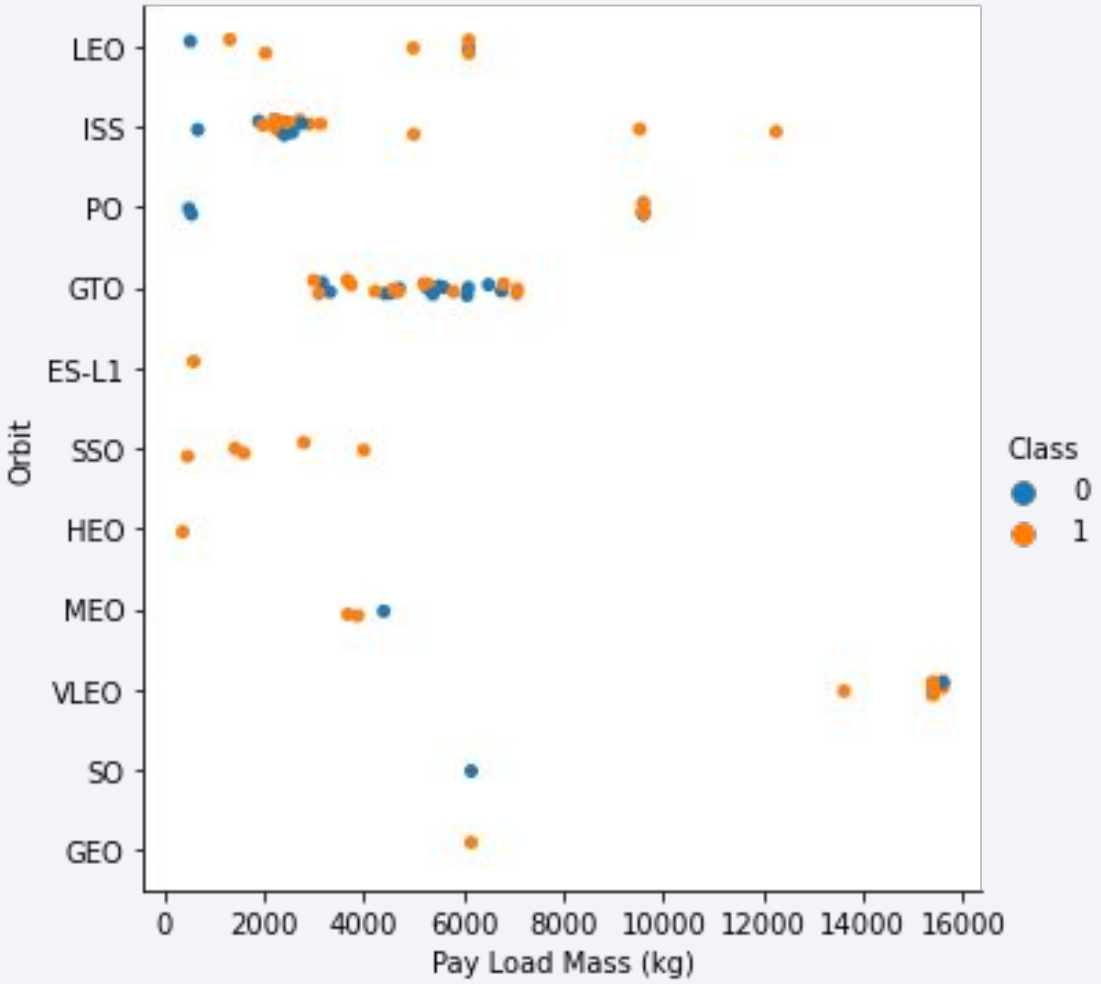
Flight Number vs. Orbit Type

- No relation in GTO orbit with Flight Number (Still have a number of failure despite higher flight number)
- LEO orbit seems to have relationship with Flight Number in terms of success rate (more flight number, higher success rate)



Payload vs. Orbit Type

- Heavy payloads has negative influence on GTO, causing more failure
- Positive influence on LEO, ISS and PO orbits. Heavier, higher success rate



Launch Success Yearly Trend

- Success rate has increased rapidly since 2015!
- Rapid development of advanced technology
- Implies the increase interest of the society towards space exploration

All Launch Site Names

Display the names of the unique launch sites in the space mission

```
In [5]: %sql SELECT DISTINCT(launch_site) FROM SPACEXTBL
```

```
 * ibm_db_sa://yxz94692:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/BLUDB
Done.
```

Out[5]:

launch_site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

4 launched sites involved:

- CCAFS LC-40
- CCAFS SLC-40
- KSC LC-39A
- VAFB SLC-4E

Launch Site Names Begin with 'CCA'

Display 5 records where launch sites begin with the string 'CCA'

```
In [6]: %sql SELECT * FROM SPACEXTBL WHERE launch_site LIKE 'CCA%' LIMIT 5
```

```
 * ibm_db_sa://yxz94692:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/BLUDB
Done.
```

Out[6]:

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- Showing records that are launched at CCA

Total Payload Mass

Display the total payload mass carried by boosters launched by NASA (CRS)

```
In [21]: %sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE CUSTOMER = 'NASA (CRS)'
```

```
 * ibm_db_sa://yxz94692:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/BLUDB
Done.
```

Out[21]:

1
45596

- Total payload mass carried by boosters launched by NASA (CRS) 45596 kg!
- **Wow! NASA invested a lot of effort and resources in space exploration!**

Average Payload Mass by F9 v1.1

Display average payload mass carried by booster version F9 v1.1

```
In [23]: %sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE BOOSTER_VERSION LIKE 'F9 v1.1%'
 * ibm_db_sa://yxz94692:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/BLUDB
Done.
```

Out[23]:

1
2534

- Average payload mass carried by booster version F9 v1.1 is 2534 kg!
- **Good booster? Maybe!**
- **Widely used? Maybe!**

First Successful Ground Landing Date

List the date when the first successful landing outcome in ground pad was acheived.

Hint:Use min function

In [24]:
```
%sql SELECT MIN(DATE) FROM SPACEXTBL
```

```
 * ibm_db_sa://yxz94692:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/BLUDB
Done.
```

Out[24]:

1
2010-06-04

- First successful landing outcome on ground pad was in 2010 June 4.
- Seeing the number of launches increased over the years, and the effort of our scientists in space exploration since 2010!

Successful Drone Ship Landing with Payload between 4000 and 6000

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
In [34]: %sql SELECT * FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000 AND landing__outcome LIKE '%Success%'
```

```
 * ibm_db_sa://yxz94692:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/BLUDB
Done.
```

Out[34]:

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2016-05-06	05:21:00	F9 FT B1022	CCAFS LC-40	JCSAT-14	4696	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
2016-08-14	05:26:00	F9 FT B1026	CCAFS LC-40	JCSAT-16	4600	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
2017-03-30	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
2017-05-01	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)
2017-09-07	14:00:00	F9 B4 B1040.1	KSC LC-39A	Boeing X-37B OTV-5	4990	LEO	U.S. Air Force	Success	Success (ground pad)
2017-10-11	22:53:00	F9 FT B1031.2	KSC LC-39A	SES-11 / EchoStar 105	5200	GTO	SES EchoStar	Success	Success (drone ship)

- There are a lot of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000! All from F9!

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

In [40]:
```
%%sql
SELECT(SELECT COUNT(*) FROM SPACEXTBL WHERE mission_outcome LIKE '%Success%') AS Success,
      (SELECT COUNT(*) FROM SPACEXTBL WHERE mission_outcome LIKE '%Failure%') AS Failure
FROM SPACEXTBL LIMIT 1
```

```
 * ibm_db_sa://yxz94692:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/BLUDB
Done.
```

Out[40]:

success	failure
100	1

- 100 success and only 1 failure mission outcomes! Failure is less than 1%
- This show that the advancement of technology and science, as well as the brilliant minds of our great scientists

Boosters Carried Maximum Payload

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

In [42]:
```
%sql SELECT booster_version FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ IN (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

```
 * ibm_db_sa://yxz94692:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/BLUDB
Done.
```

Out[42]:

booster_version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

- Best booster is F9 carrying lots of Maximum payload mass? Maybe!

2015 Launch Records

List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015 ¶

```
In [44]: %sql SELECT booster_version, launch_site FROM SPACEXTBL WHERE LANDING__OUTCOME LIKE '%Fail%' AND YEAR(DATE) = 2015
 * ibm_db_sa://yxz94692:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/BLUDB
Done.
```

Out[44]:

booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

- Only 2 launches failed in landing in 2015! Thanks to the advancement of technologies and science development! A great improvement, lesser lost

Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

In [47]:
```
%%sql

SELECT LANDING__OUTCOME,COUNT(LANDING__OUTCOME) AS Countland
FROM SPACEXTBL
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY LANDING__OUTCOME
ORDER BY Countland DESC
```

```
 * ibm_db_sa://yxz94692:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu01qde00.databases.appdomain.cloud:31321/BLUDB
Done.
```

Out[47]:

landing__outcome	countland
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

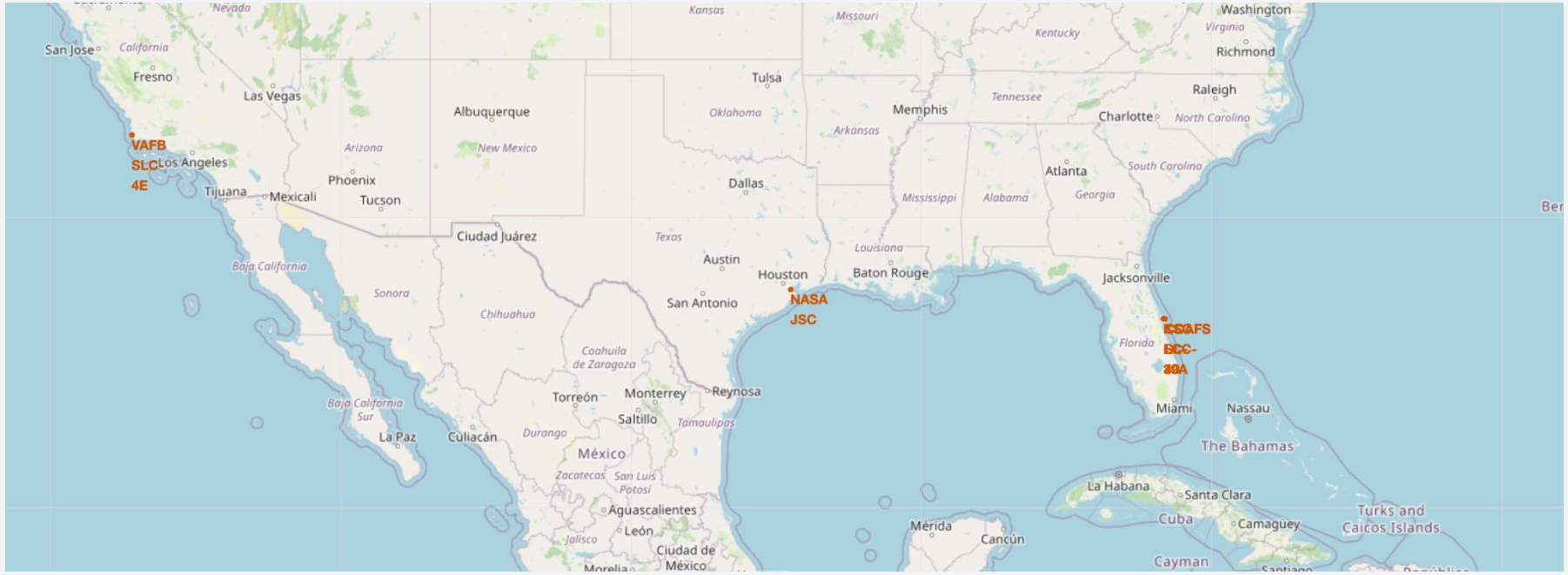
- Most of the launches in between 2010-06-04 and 2017-03-20 are ‘No attempt’, insufficient resources or costs? probably! An issue that authorities should bring attention to!

Section 4

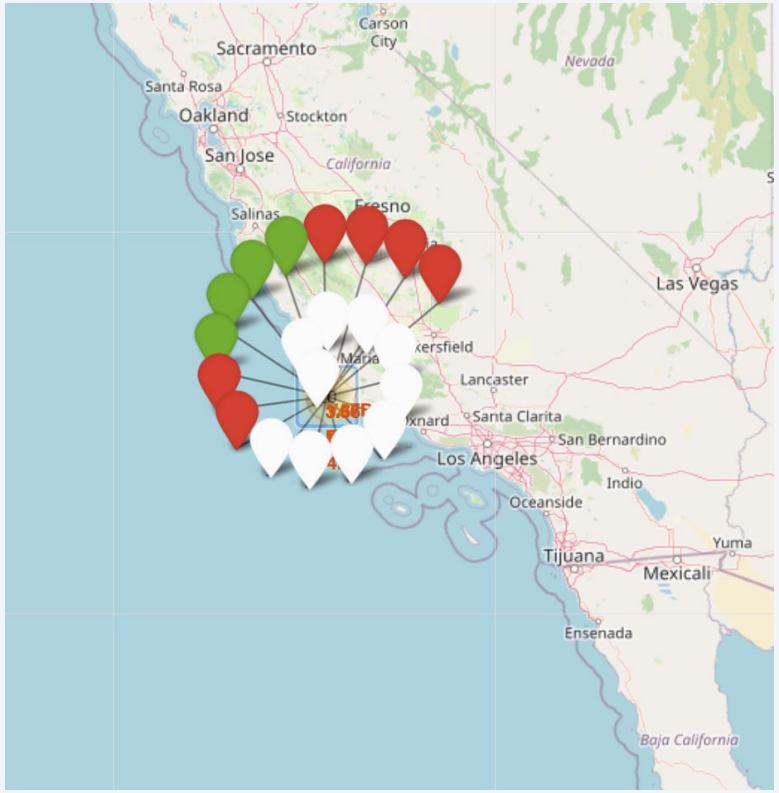
Launch Sites Proximities Analysis

Launch Site location on Map

- All launch site involved are marked on the location
- Observed that they are generally close to the coast line … hmm why? a problem worth to investigate on!



Success or Failure? Good spot for launch?

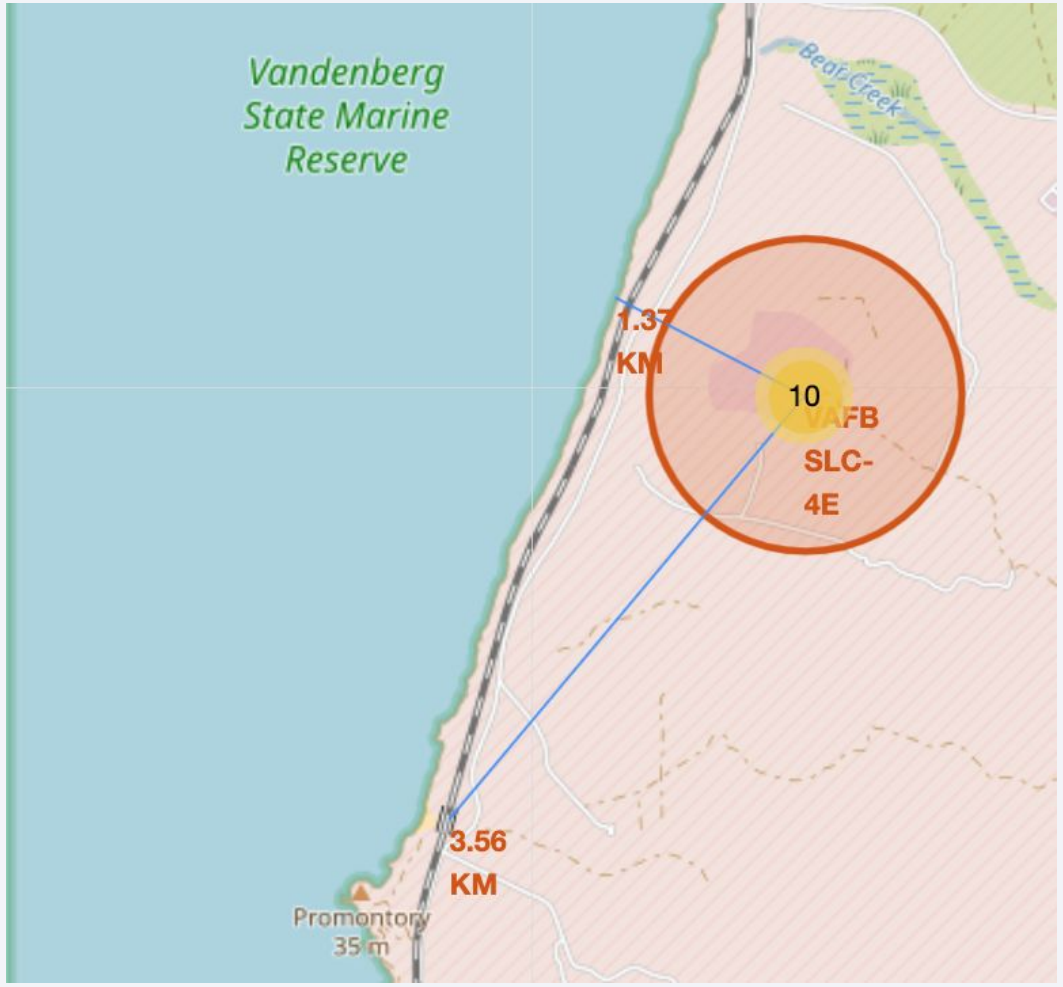


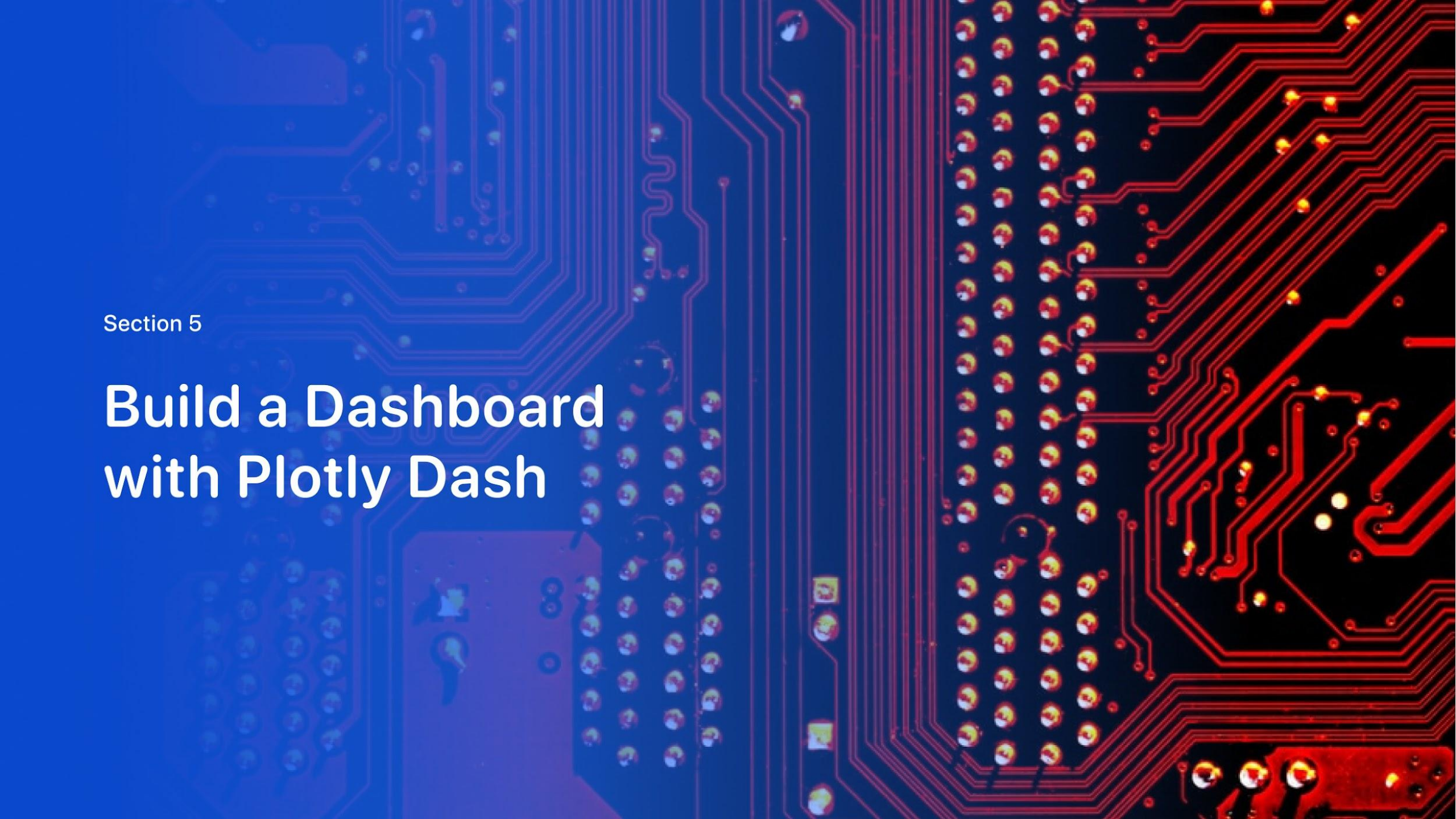
- Red → Failure
- Green → Success launch
- White → No Attempt/None

Most of them doesn’t not have attempt :(

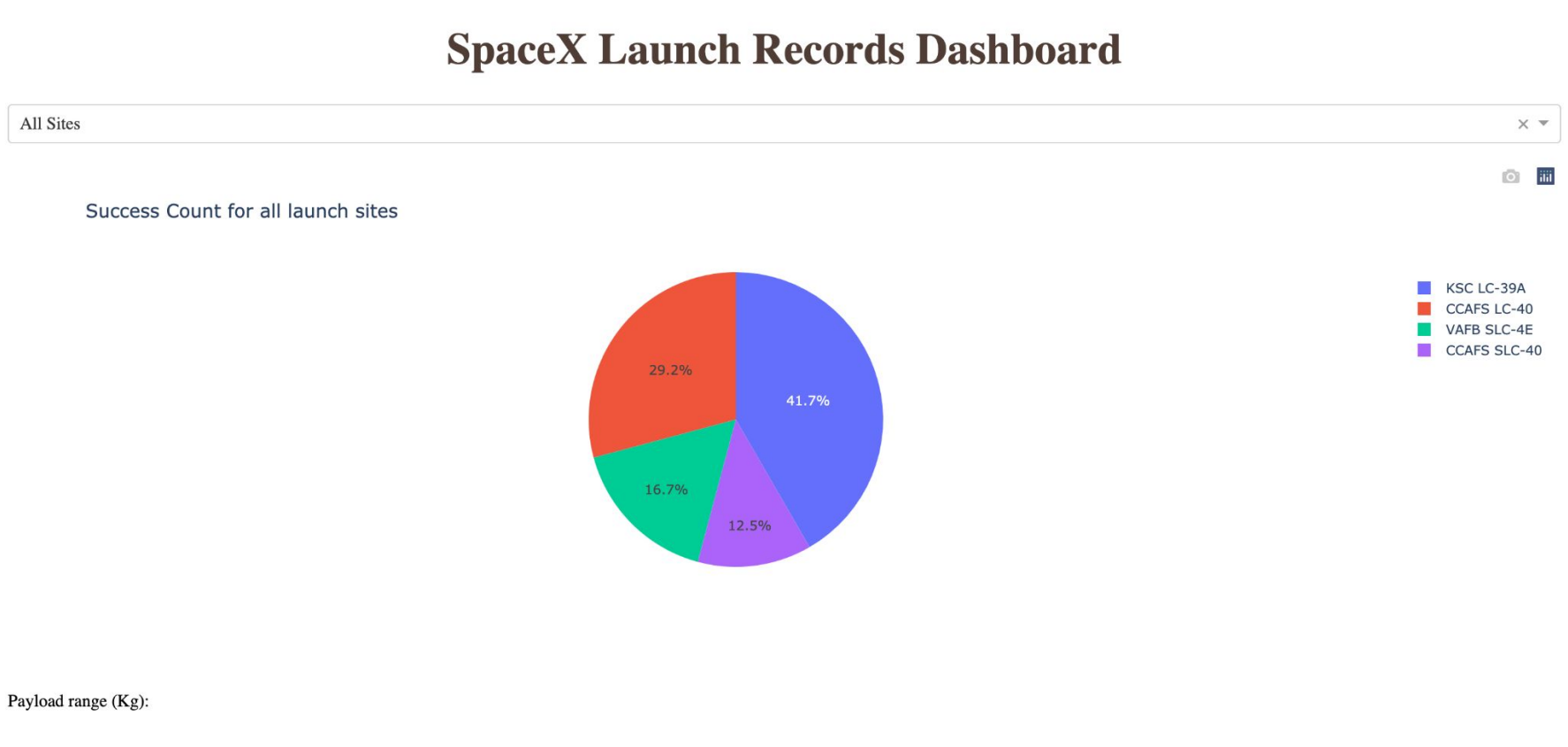
Close to proximities?

- 1.37 KM to cost and 3.56 KM to railway!
- Launch site are better at coast line? Is there any other specific reason behind building launch site near the cost?
- Is it for easier observation when landing in the sea?
- Maybe!



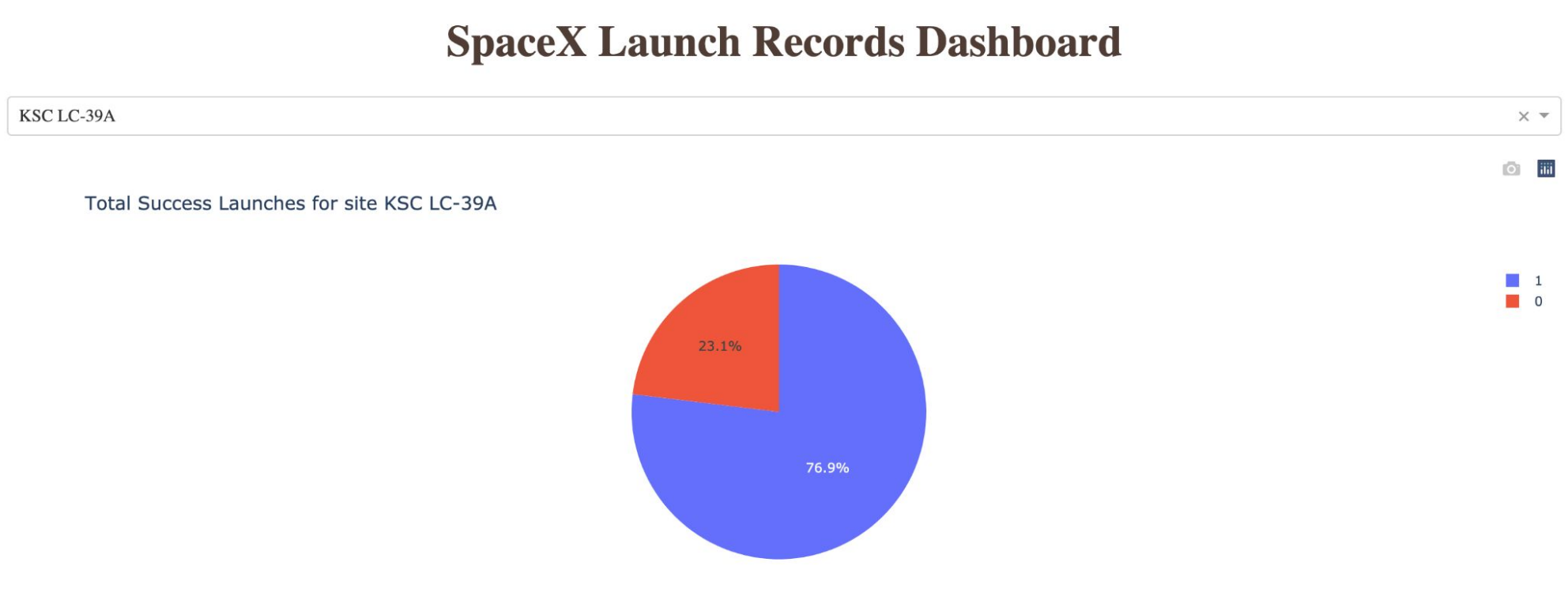

Section 5
Build a Dashboard with Plotly Dash

SpaceX Launch Record Dashboard



- KSC LC-39A has the highest success rate! Best spot maybe!

How good is KSC LC-39A at launching?



- 77% of the launches are successful! Wow! impressive results!

<Dashboard Screenshot 3>

- FT and B4 booster for high payload mass range! however, more failures is happening! Oh no!

Section 6

Predictive Analysis (Classification)

Classification Accuracy

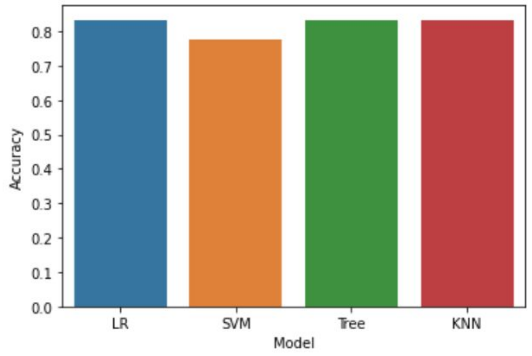
- SVM has the lowest accuracy, not suitable for prediction success/failure launching
- Other models has equal performance of accuracy up to 83.3%!

```
In [49]: score

Out[49]: [0.8333333333333334,
          0.7777777777777778,
          0.8333333333333334,
          0.8333333333333334]

In [50]: model = ["LR","SVM","Tree","KNN"]
         acc = pd.DataFrame(list(zip(model, score)),columns =['Model', 'Accuracy'])
         acc
```

Out[50]:

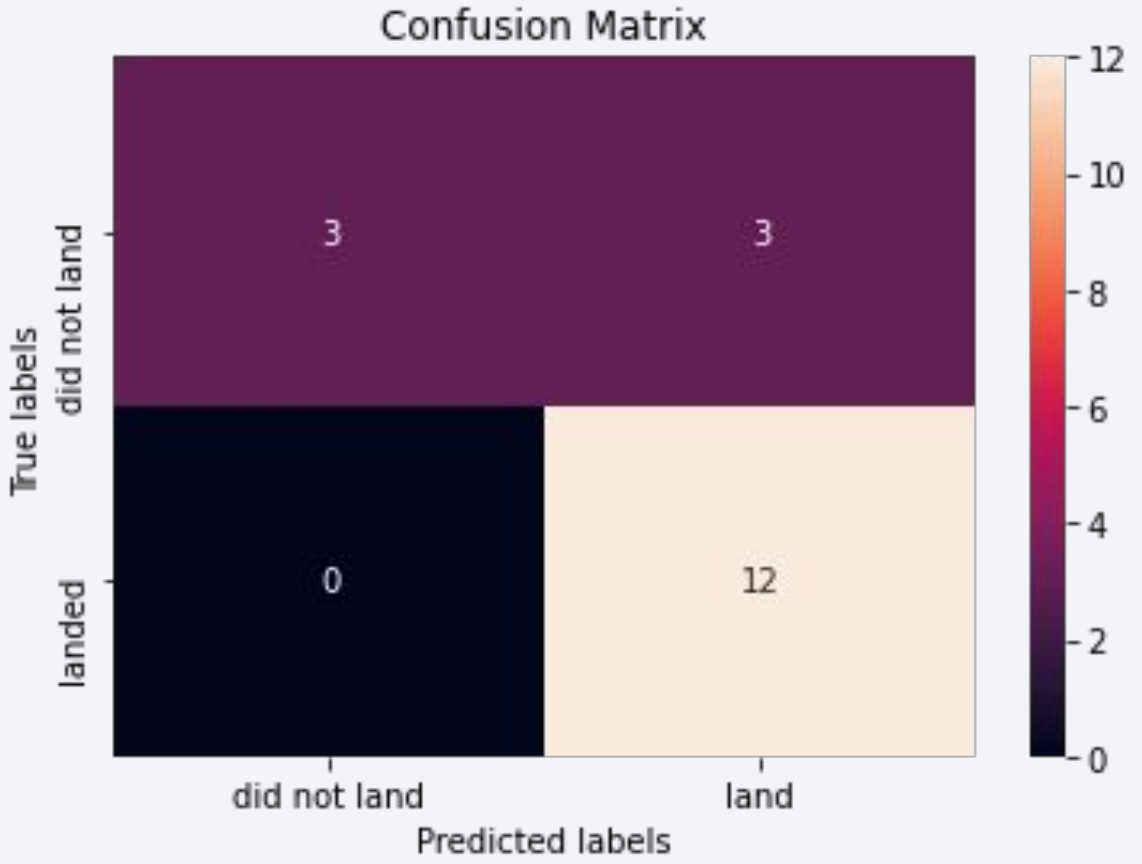
	Model	Accuracy
0	LR	0.833333
1	SVM	0.777778
2	Tree	0.833333
3	KNN	0.833333

```
In [51]: sns.barplot(x="Model",y="Accuracy",data=acc)

Out[51]: <AxesSubplot:xlabel='Model', ylabel='Accuracy'>
```



Confusion Matrix

- In this case study, False Positive (FP) is the worst case scenario as the launch that will fail is predicted success, which will cause a huge loss in effort and resources
- The FP is considerably high with 3 FP in a small test set
- Model should be refined and adjusted!



Conclusions

- SpaceX can consider launching to Orbit ES-L1 GEO, HEO and SSO that has higher success rate
- F9 booster can be used if SpaceX has a high payload mass
- KSC LC-39A has the highest success rate, SpaceX could take advantage of this and have launching there
- Launch sites are generally close to the coast line for easier observation
- SVM is not suitable for predictions of success / failure of launches
- FP is the worst case scenario, model should be refined to minimize FP

Appendix

- All assets provided by IBM Data Science professional certificate course

Thank you!